Learning Decoupled Representations for Human Pose Forecasting

Behnam Parsaeifard[1,2,*] Saeed Saadatnejad[2,*] Yuejiang Liu[2] Taylor Mordan[2] Alexandre Alahi[2]

[1]University of Basel, Switzerland [2]Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland

behnam.parsaeifard@unibas.ch saeed.saadatnejad@epfl.ch

Abstract

Human pose forecasting involves complex spatiotemporal interactions between body parts (e.g., arms, legs, spine). State-of-the-art approaches use Long Short-Term Memories (LSTMs) or Variational AutoEncoders (VAEs) to solve the problem. Yet, they do not effectively predict human motions when both global trajectory and local pose movements exist. We propose to learn decoupled representations for the global and local pose forecasting tasks. We also show that it is better to stop the prediction when the uncertainty in human motion increases. Our forecasting model outperforms all existing methods on the pose forecasting benchmark to date by over 20%. The code is available online [†].

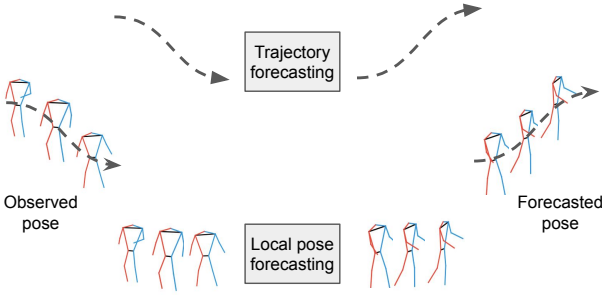


Figure 1: Decoupling the human pose into a trajectory and local pose. The dashed arrows indicate the trajectory of the human.

1. Introduction

Human pose forecasting is defined as predicting future human keypoints' locations –the body parts (e.g., legs, arms, spine)– given a sequence of observed ones. It has attracted more attention in recent years due to its critical applications in self-driving cars [34], robotics [42, 12], and healthcare [25, 47, 44, 10]. For example, in self-driving cars, it is very important to predict the location of pedestrians to avoid accidents [28]. Furthermore, the body pose of pedestrians often provide useful information about whether or not they intend to cross the street [40]. Unfortunately, the high uncertainty in this problem makes it challenging such that even we, humans, are often not able to exactly predict the next motions. In this work, we want to learn a representation of human pose dynamics to effectively predict plausible motions and potentially stop predicting when the uncertainty is too high.

The human pose forecasting task can be decoupled into a global (coarse) trajectory forecasting task and a local (fine-grained) pose forecasting one. At the coarse level, the large-scale movements of humans with respect to the camera are modeled. However, at the fine-grained level, all the detailed local movements of different keypoints are modeled. Pioneering works showed promising results for trajectory forecasting [3, 20] and local pose forecasting *i.e.*, excluding the global trajectory movements [36, 5]. They used Long Short-Term Memories (LSTMs) because of their ability to capture temporal dependencies or Variational Autoencoders (VAEs) because of their ability in generating a new pose considering the non-deterministic task. While they achieved outstanding results for each of these separate tasks, they have limited performance to predict the human pose dynamics when both trajectory and local pose move.

Considering the complexity of this task, we propose to decompose it into trajectory forecasting and local pose forecasting tasks (see Figure 1). When a person moves, their global coordinates and the local coordinates of their keypoints (with respect to their trajectory) change in different ways and this distinction helps us exploit different approaches for both.

We propose an LSTM encoder-decoder network for trajectory forecasting and a VAE-encoder-decoder to solve this local pose forecasting task. Moreover, if the network is not confident about the future, it stops predicting and takes the last prediction. We show that using this approach results in a

*Equal contribution, order chosen alphabetically

[†]https://github.com/vita-epfl/decoupled-pose-prediction.git

significant improvement of the quality of the predicted pose both visually and numerically in the evaluation metrics.

2. Related works

2.1. Trajectory Forecasting

Trajectory forecasting refers to the task of predicting the future trajectories of humans, *i.e.*, the future positions of them over time (in XY coordinates) based on previous locations. In the pioneering work [21], the authors presented the social force model for trajectory forecasting, in which attractive and repulsive forces are introduced to impose physical constraints such as navigating the person toward their goal while forcing them to keep distance from other people in the scene. This and similar hand-crafted models [48, 33, 37, 4, 39, 41, 15] rely on small datasets at the cost of limited prediction accuracy.

The data-driven and deep learning methods have been proposed to increase the prediction accuracy [17, 20, 31, 3, 23, 7, 30, 29, 32] of the hand-crafted models. These approaches rely on large datasets [28]. Due to the sequential nature of the human trajectories, a history of 2D or 3D coordinates, recurrent neural network (RNNs) and their variants can be used. Social LSTM [3] proposed a Long Short-Term Memory (LSTM) network with a social pooling module modeling the interaction between the humans. Others studied different deep learning alternatives such as feed-forward networks [22], graph attention network [23] and a deep generative adversarial network (GAN) coupled with a recurrent neural network [20]. In this paper, we use an encoder-decoder LSTM network for our trajectory forecasting task.

2.2. Pose Forecasting

The trajectory forecasting only provides coarse information about the future but we can go one step further and predict the bounding boxes around the human to catch the size of them too [8]. This can be extended further to predicting the future body keypoints over time based on previous observations.

Existing works on pose forecasting mostly ignore the global motion of the human and only predict the changes in keypoints locations with respect to the center of human with the global motion excluded [18, 24, 36, 11, 19, 13, 35, 14, 45, 5, 49]. RNNs, capable of capturing the temporal dependencies in sequential data, have been widely used for the problem of local pose forecasting [18, 24, 36, 11, 19, 13, 35]. Combining high-level spatio-temporal graphs with RNNs [24], sequence-to-sequence architecture with residual connections [36] and forecasting human dynamics from static images [11] are used for pose forecasting. There are some other ideas for pose forecasting including combining a 3-layer LSTM and a dropout autoencoder [19], hierarchical RNN [13], attention-based feed-forward network [35], completing an incomplete pose and designing a graph convolutional network [16], and including context using a graph attention [14]. There are also some probabilistic approaches based on GANs [45] and VAEs [5, 49] considering their strengths in generation and learning representations. In [5], the authors proposed an encoder-decoder network with a conditional VAE for pose forecasting. In [49], the authors proposed Motion Transformation VAE (MT-VAE) to generate multiple plausible pose from the same input.

However, none of those methods solved the pose forecasting problem when both the trajectory and local pose change. There are still some works in the literature that predicted pose when they move globally [2, 1, 9, 34, 46]. Some of them proposed the goal-directed human motion forecasting by incorporating the context of the scene in the prediction [9] or synthesizing human motion between two points given the inputs of start and end positions [46]. Closer to ours, SC-MPF [1] and TRiPOD [2] predicted the trajectory and local pose dynamics as a single task by considering various human-human, human-objects, and human-scene interactions. In another approach, the human motion was split into the global and local dynamics [34] in a deterministic way. In this work, we decouple the human pose into a trajectory and a local pose and propose a generative approach based on VAE to learn a representation for the local pose dynamics. Furthermore, we show that our approach outperforms those baselines significantly in evaluation metrics.

3. DeRPoF: Decoupled Representations for Pose Forecasting

In this section, we describe our method to decouple representations for pose forecasting. Our model consists of two networks for trajectory and local pose forecastings and the general concept of decoupling is depicted in Figure 1.

3.1. Formulation

Given a sequence of T_{obs} observed pose $\{\boldsymbol{P}_1, \boldsymbol{P}_2, \dots, \boldsymbol{P}_{T_{obs}}\}$, where $\boldsymbol{P}_t = \{\boldsymbol{P}_t^i\}_{i=1:d}$ is the location of d keypoints at time step t, we aim at predicting the next T_{future} future pose, *i.e.*, $\{\boldsymbol{P}_{T_{obs}+1}, \boldsymbol{P}_{T_{obs}+2}, \dots, \boldsymbol{P}_{T_{obs}+T_{future}}\}$. We split the keypoints $\boldsymbol{P}_t^i$ at each time t into two parts:

$$\boldsymbol{P}_t^i = \boldsymbol{R}_t + \boldsymbol{r}_t^i, \tag{1}$$

where $\boldsymbol{R}_t$ is the location of the center of the human with respect to the camera and $\boldsymbol{r}_t = \{\boldsymbol{r}_t^i\}_{i=1:d}$ is the locations of d keypoints with respect to $\boldsymbol{R}_t$. The trajectory $\{\boldsymbol{R}_t\}_{t=1:T_{obs}}$ represents the global movements of the human with respect to the camera and the local pose sequence $\{\boldsymbol{r}_t\}_{t=1:T_{obs}}$ indicates the fine-grained movements of keypoints with respect to the trajectory.

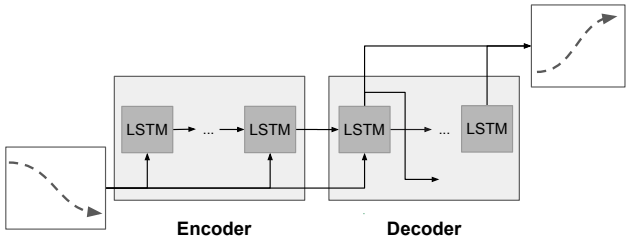


Figure 2: The LSTM encoder-decoder network for trajectory forecasting. The dashed arrows indicate the trajectory of the human.

Similar to previous works [36, 1], instead of locations, we use velocities which are more effective in the learning process. The velocity of the pose at time t is simply $\{\dot{\boldsymbol{P}}_t^i\} = \boldsymbol{P}_t - \boldsymbol{P}_{t-1}$ and its i-th element ($i \in 1, \ldots, d$) can be written as:

$$\dot{\boldsymbol{P}}_t^i = \dot{\boldsymbol{R}}_t + \dot{\boldsymbol{r}}_t^i \tag{2}$$

where $\dot{\boldsymbol{R}}_t = \boldsymbol{R}_t - \boldsymbol{R}_{t-1}$, $\dot{\boldsymbol{r}}_t = \{\dot{\boldsymbol{r}}_t^i\}_{i=1:d} = \{\boldsymbol{r}_t^i - \boldsymbol{r}_{t-1}^i\}_{i=1:d}$.

3.2. Network Architecture

Our method comprises two networks: an LSTM encoder-decoder for trajectory forecasting and a VAE encoder-decoder for the local pose forecasting.

3.2.1 Trajectory Forecasting

Figure 2 shows this part of the network. We use an LSTM encoder-decoder network similar to [8] for the trajectory forecasting. The encoder LSTM at time t takes in as input the observed velocities $\{\dot{\boldsymbol{R}}_1, \ldots, \dot{\boldsymbol{R}}_t\}$ and outputs the updated hidden state, *i.e.*:

$$\boldsymbol{h}_t = LSTM_{enc}^{traj}(\boldsymbol{h}_{t-1}, \dot{\boldsymbol{R}}_t) \tag{3}$$

where $LSTM_{enc}^{traj}$ is the encoder LSTM. The encoder extracts the important features of the input and encodes them in the hidden state.
After encoding the input, the decoder LSTM takes as input the last observed velocity as well as the last hidden state of the encoder as its initial hidden state and outputs the predicted hidden state, *i.e.*:

$$\hat{\boldsymbol{h}}_{t+1} = LSTM_{dec}^{traj}(\boldsymbol{h}_t, \dot{\boldsymbol{R}}_t) \tag{4}$$

where $LSTM_{dec}^{traj}$ is the decoder LSTM. A fully connected layer is finally used to predict the next future velocity, *i.e.*:

$$\hat{\dot{\boldsymbol{R}}}_{t+1} = \phi(\hat{\boldsymbol{h}}_{t+1}) \tag{5}$$

where ϕ is the fully connected layer. We can predict the next velocities from Equations 4 and 5 if we use as input in Equation 4 the previous predicted hidden state and velocity, *i.e.*, for a later time $t + t'$ we write:

$$\begin{aligned} \hat{\boldsymbol{h}}_{t+t'} &= LSTM_{dec}^{traj}(\hat{\boldsymbol{h}}_{t+t'-1}, \hat{\dot{\boldsymbol{R}}}_{t+t'-1}) \\ \hat{\dot{\boldsymbol{R}}}_{t+t'} &= \phi(\hat{\boldsymbol{h}}_{t+t'}) \\ \hat{\boldsymbol{R}}_{t+t'} &= \hat{\boldsymbol{R}}_{t+t'-1} + \hat{\dot{\boldsymbol{R}}}_{t+t'} \end{aligned} \tag{6}$$

The velocities are used to iteratively compute the future trajectory using the above equation.

3.2.2 Local Pose Forecasting

A common VAE consists of an encoder that maps the input to a distribution in a latent space and a decoder that samples from that distribution and tries to regenerate the input. The optimal parameters of the network are found by minimizing the reconstruction loss, defined usually as the Mean Squared Error (MSE) between the prediction and the input, as well as the KL Divergence between the latent space distribution and a standard Gaussian distribution [27]. In our task of local pose forecasting, we use a slightly different VAE in which the network does not regenerate the input but tries to generate the future ground truths. In this case, the loss is defined as the MSE between the predictions and the ground truths in addition to the KL divergence.

We show the architecture of the network in Figure 3. The encoder LSTM, $LSTM_{enc}^{local}$, at time t takes in as input the sequence of the observed local pose velocities, $\{\dot{\boldsymbol{r}}_1, \ldots, \dot{\boldsymbol{r}}_t\}$, and outputs the updated hidden state, *i.e.*:

$$\boldsymbol{h}_t^{'} = LSTM_{enc}^{local}(\boldsymbol{h}_{t-1}^{'}, \dot{\boldsymbol{r}}_t) \tag{7}$$

Two fully connected layers take as input the hidden state and output the mean $\boldsymbol{\mu}$ and the covariance $\boldsymbol{\sigma}$, *i.e.*:

$$\begin{aligned} \boldsymbol{\mu} &= \phi_{\boldsymbol{\mu}}(\boldsymbol{h}_t^{'}) \\ \boldsymbol{\sigma} &= \phi_{\boldsymbol{\sigma}}(\boldsymbol{h}_t^{'}) \\ \boldsymbol{z} &= \boldsymbol{\mu} + \boldsymbol{\sigma} * \boldsymbol{\xi} \end{aligned} \tag{8}$$

where $\phi_{\boldsymbol{\mu}}$ and $\phi_{\boldsymbol{\sigma}}$ are the fully connected layers and $\boldsymbol{\xi}$ is a random variable sampled from a multivariate standard Gaussian using the reparameterization trick [27]. Having calculated $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, we set the initial hidden state of the decoder to a fully connected layer applied on $\boldsymbol{z}$.

The decoder LSTM, $LSTM_{dec}^{local}$, takes as input the last observed local pose velocity as well as the hidden state and outputs the predicted hidden state, *i.e.*:

$$\hat{\boldsymbol{h}}_{t+1}^{'} = LSTM_{dec}^{local}(\boldsymbol{h}_t^{'}, \dot{\boldsymbol{r}}_t) \tag{9}$$

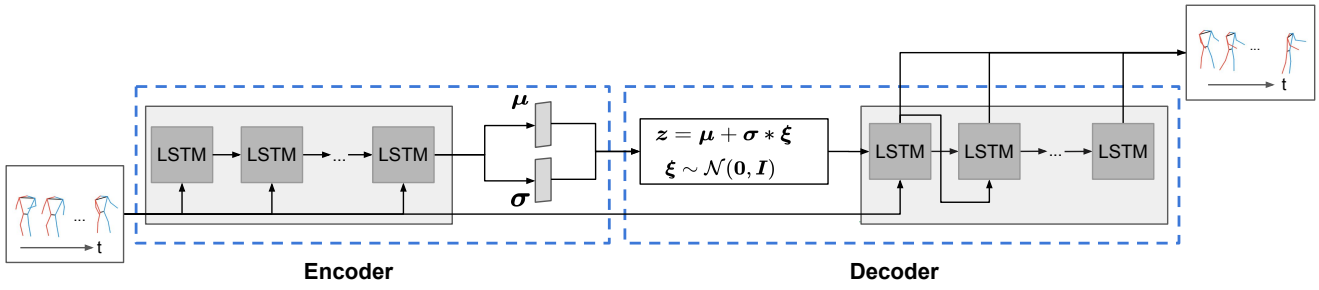


Figure 3: The VAE encoder-decoder network for the local pose forecasting.

We use a fully connected layer, $\phi^{'}$, to predict the next future local pose velocity, *i.e.*:

$$\hat{\dot{\boldsymbol{r}}}_{t+1} = \phi^{'}(\hat{\boldsymbol{h}}^{'}_{t+1}) \tag{10}$$

The next velocities are predicted from Equations 9 and 10 by using as input in Equation 9 the previous predicted hidden state and velocity, *i.e.*, for a later time $t+t'$ we write:

$$\begin{aligned} \hat{\boldsymbol{h}}^{'}_{t+t'} &= LSTM^{local}_{dec}(\hat{\boldsymbol{h}}^{'}_{t+t'-1}, \hat{\dot{\boldsymbol{r}}}_{t+t'-1}) \\ \hat{\dot{\boldsymbol{r}}}_{t+t'} &= \phi^{'}(\hat{\boldsymbol{h}}^{'}_{t+t'}) \\ \hat{\boldsymbol{r}}_{t+t'} &= \hat{\boldsymbol{r}}_{t+t'-1} + \hat{\dot{\boldsymbol{r}}}_{t+t'} \end{aligned} \tag{11}$$

These velocities are used to iteratively compute the future local pose in the above equation.

3.3. Training

The objective is to achieve realistic and accurate pose close to the ground truth pose while keeping the trajectory and the local pose decoupled. Therefore, the loss is derived as follows:

$$\mathcal{L} = \sum_{t=T_{obs+1}}^{T_{obs}+T_{future}} \Big(||\hat{\ddot{\boldsymbol{R}}}_t - \ddot{\boldsymbol{R}}_t||^2 + \lambda_l ||\hat{\dot{\boldsymbol{r}}}_t - \dot{\boldsymbol{r}}_t||^2 \Big) + \lambda_k KL\Big(P(\boldsymbol{z}) \,||\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \Big) \tag{12}$$

where $p(\boldsymbol{z})$ is the probability distribution of $\boldsymbol{z}$ and λ_l and λ_k are two hyperparameters to control the contribution of each term in the final loss. The first term in Equation 12 penalizes the trajectory forecasting and the second and third terms penalize realistic local pose forecasting.

We also consider the uncertainty in the pose forecasting. Once the model is uncertain, it stops predicting and outputs the last prediction. We call the model at $t > t_c$, where t_c is the threshold, uncertain if the model is penalized less by predicting up to time t_c and using the last prediction as the future predictions for times $t > t_c$. In the beginning of the training, t_c starts from 0 but it increases after some epochs of training and reaches a value smaller or equal to the maximum prediction length.

4. Experiment

In this section, we first introduce datasets, evaluation metrics, baselines, and experiment details. To be fair in the results, we follow the same procedure mentioned in the SoMoF challenge [2].

4.1. Datasets

- **3DPW**[43] 3D Pose in the Wild is an accurate dataset with 3D pose containing various indoor and outdoor scenarios such as walking, arguing, phoning, doing sport, etc. We use a subset of this dataset used as benchmarking in the SoMoF challenge. This subset contains 220 sequences in the train, 36 sequences in the validation, and 85 sequences in the test set. The pose consist of 13 body keypoints in 3D including the neck, shoulders, elbows, wrists, knees, hips, and ankles.

- **PoseTrack**[6] is another dataset containing various indoor as well as outdoor scenarios originally purposed for the problem of pose estimation. PoseTrack dataset contains human keypoints which are partly occluded and invisible. Similarly, we use the official PoseTrack dataset in the SoMoF challenge. This dataset has 306, 104, and 106 sequences in the train, validation, and test respectively. The pose consist of 14 body keypoints in 2D including the head, neck, shoulders, elbows, wrists, knees, hips, and ankles.

We have taken 16 frames (640 ms in PoseTrack and 1030 ms in 3DPW) as the observation and predicted up to the next 14 frames (560 ms in PoseTrack and 900 ms in 3DPW).

4.2. Evaluation Metrics

- **VIM**: Visibility Ignored Metric is the average Euclidean distance between the location of the predicted keypoints and the location of the keypoints in the ground truth [2].

- **VAM**: Visibility Aware Metric is the same as VIM if the visibility of the keypoints is predicted correctly

otherwise a penalty B is imposed [2]. This value is considered as 200 in our experiments.

4.3. Baselines

- **Zero velocity**: A simple baseline to keep the last observed pose as the prediction for future frames. It has been shown that this baseline is a hard-to-beat baseline that has outperformed a large number of models [36, 18].
- **Nearest neighbour**: This is a simple baseline in which using Euclidean distance, we compare the normalized input sequence in the test set with the normalized input sequences in the train set and choose the closest one. Then, the prediction will be the future sequence of that selected sequence. This baseline has been compared with in some other works too [50].
- **TRiPOD [2]**: This is a recent work on human pose forecasting which takes into account various interactions between the humans and other humans and objects in the scene using graph attention networks.
- **SC-MPF [1]**: It is another recent method for pose forecasting. The human-to-human interaction is included in this model using social pooling.
- Other baselines (PF-RNN + S-LSTM, MoAtt + S-LSTM, PF-RNN + S-GAN, MoAtt + S-GAN, PF-RNN + ST-GAT and MoAtt + ST-GAT): In these models, the human pose forecasting is achieved by combining the results of the trajectory forecasting (MoAtt [35] or PF-RNN [36]) and the local pose forecasting (ST-GAT [23], S-LSTM [3] or S-GAN [20]).

4.4. Implementation Details

The dimension of the hidden states of all the LSTMs in the model is 64. The dimension of the latent space in the VAE is 32. We have taken λ_l and λ_k in Equation 12 to be 1 and 0.01 respectively. We have implemented our model in Pytorch [38] and trained it using the Adam optimizer [26]. The learning rate starts at 0.001 and is updated during the training by an adaptive scheduler. All the models are trained for 1000 epochs on an NVIDIA GTX-2080-Ti GPU.

4.5. Results and Discussions

We evaluate our model on the unseen test dataset with the evaluation metrics introduced in Section 4.2. The quantitative numbers for 3DPW and Posetrack datasets are reported in Table 1 and Table 2, respectively. Our model (DeRPoF) outperforms all the previous baselines including two-stage and one-stage predictions in both datasets. Furthermore, it shows that given the high uncertainty in human pose forecasting, sometimes not predicting (Zero velocity) is better than predicting. We used the numbers reported in the challenge and unfortunately, we could not compare with [34] as their source code is not available and our re-implementation shows non-realistic results.

To study the effect of each module in the final outcome, we do an ablation study:

- **w/o Early stop** In this model, our network predicts all the future frames regardless of the uncertainty of the predictions.
- **w/o Decoupling** In this model, we remove the decoupling from the model, *i.e.,* the human motion is not split into a global trajectory and local pose. In this baseline, the pose is predicted using the VAE-encoder-decoder.
- **w/o VAE, Decoupling** In this model, we remove both the VAE and the decoupling, *i.e.* the human motion is predicted without splitting it into global and local motion using an encoder-decoder LSTM-based network.

The quantitative results of all these modes are reported in Table 3. This experiment is conducted on the 3DPW dataset. As it shows, each part improves the performance and all of them are required to capture an accurate human motion.

To better show it, the qualitative results of these models are depicted in Figure 4. The observed pose for $t = 2, 4, 6, 8, 10, 12, 14, 16$ are on the left and the predicted pose for $t = 17, 19, 21, 23, 25, 27, 29$ are on the right for two different scenarios (a) and (b). The keypoints move more naturally in the two top rows (DeRPoF and w/o Early stop) of each scenario while the keypoints do not move in a real manner when the decoupling is omitted. This is in accordance with our assumption about the necessity of learning decoupled representations to predict a realistic human motion. The predictions of Nearest neighbor and Zero velocity baselines are presented in the last two rows. We could not include the qualitative results for other baselines since their codes are not available.

Using a generative model provides us the ability to sample from the learned distribution and generate multiple examples per a given sequence of human motion. In the above analysis, we used only the mean of the learned representation but we could verify that it generates multiple distinct samples.

5. Conclusions and Future Works

We have tackled the human pose forecasting task by decomposing it into global trajectory and local pose forecasting tasks. We employed a simple LSTM encoder-decoder network for the prediction of the trajectory and proposed a VAE-encoder-decoder for the local pose forecasting. Evaluating our model on the 3DPW and PoseTrack datasets in

method	Prediction time				
	100 ms	240 ms	500 ms	640 ms	900 ms
PF-RNN [36]+S-LSTM [3]	73.82	127.23	179.07	202.78	277.55
MoAtt [35]+S-LSTM [3]	64.64	111.67	168.67	202.16	267.65
PF-RNN [36]+S-GAN [20]	83.35	138.48	182.84	204.84	291.96
MoAtt [35]+S-GAN [20]	66.36	112.18	166.48	209.53	277.85
PF-RNN [36]+ST-GAT [23]	66.95	117.77	165.99	190.52	252.23
MoAtt [35]+ST-GAT [23]	62.15	97.74	155.23	184.96	250.98
SC-MPF [1]	46.28	73.88	130.23	160.83	208.44
TRiPOD [2]	30.26	51.84	85.08	104.78	146.33
Nearest neighbour	27.34	51.68	97.75	121.40	168.27
Zero velocity	29.35	53.56	94.52	112.68	143.10
DeRPoF (ours)	**19.53**	**36.89**	**68.29**	**85.45**	**118.21**

Table 1: Comparison of VIM for 3DPW dataset. The reported numbers are in centimeter.

method	Prediction time				
	80 ms	160 ms	320 ms	400 ms	560 ms
PF-RNN [36]+S-LSTM [3]	87.02/89.44	103.22/111.11	129.21/136.43	138.77/145.72	160.95/157.04
MoAtt [35]+S-LSTM [3]	82.06/86.76	100.99/109.02	121.43/130.82	132.73/142.35	156.15/155.21
PF-RNN [36]+S-GAN [20]	84.40/87.23	99.24/106.23	130.21/131.12	130.21/139.94	150.03/150.44
MoAtt [35]+S-GAN [20]	82.45/85.82	98.76/104.13	119.38/128.97	127.98/139.07	149.53/151.45
PF-RNN [36]+ST-GAT [23]	82.06/86.76	94.25/102.61	117.70/127.87	126.71/137.87	148.65/150.80
MoAtt [35]+ST-GAT [23]	80.60/86.29	93.43/100.92	115.68/125.32	129.54/137.50	141.13/147.92
SC-MPF [1]	22.01/78.36	37.99/ 99.80	64.62/124.38	75.84/138.52	93.54/147.93
TRiPOD [2]	15.21/30.00	26.79/ 49.66	48.12/ 80.32	58.68/ 93.32	74.11/110.40
Nearest neighbour	11.75/24.62	21.35/ 42.05	41.15/ 70.95	50.99/ 82.76	66.80/ 99.91
Zero velocity	13.17/26.57	24.06/ 45.17	43.31/ 72.92	52.17/ 83.87	65.63/ 97.34
DeRPoF (ours)	**10.20/22.05**	**18.56/37.29**	**34.89/62.01**	**42.76/73.10**	**54.62/88.12**

Table 2: Comparison of VIM/VAM for PoseTrack dataset. The reported results are in pixel.

method	Prediction time				
	100 ms	240 ms	500 ms	640 ms	900 ms
DeRPoF	19.53	36.89	68.29	85.45	118.21
w/o Early stop	19.53	36.89	70.70	89.02	126.19
w/o Decoupling	19.27	36.84	71.02	89.76	127.73
w/o VAE, Decoupling	20.50	37.95	72.68	91.94	131.99

Table 3: The ablation study of VIM for 3DPW dataset. The reported numbers are in centimeter.

the SoMoF benchmarking, we have shown that our model outperforms all the baselines in the challenge.

Motivated by the success of our work in improving the accuracy of human motion forecasting in 3DPW and PoseTrack, this can be applied to other datasets especially for longer predictions and in different environments.

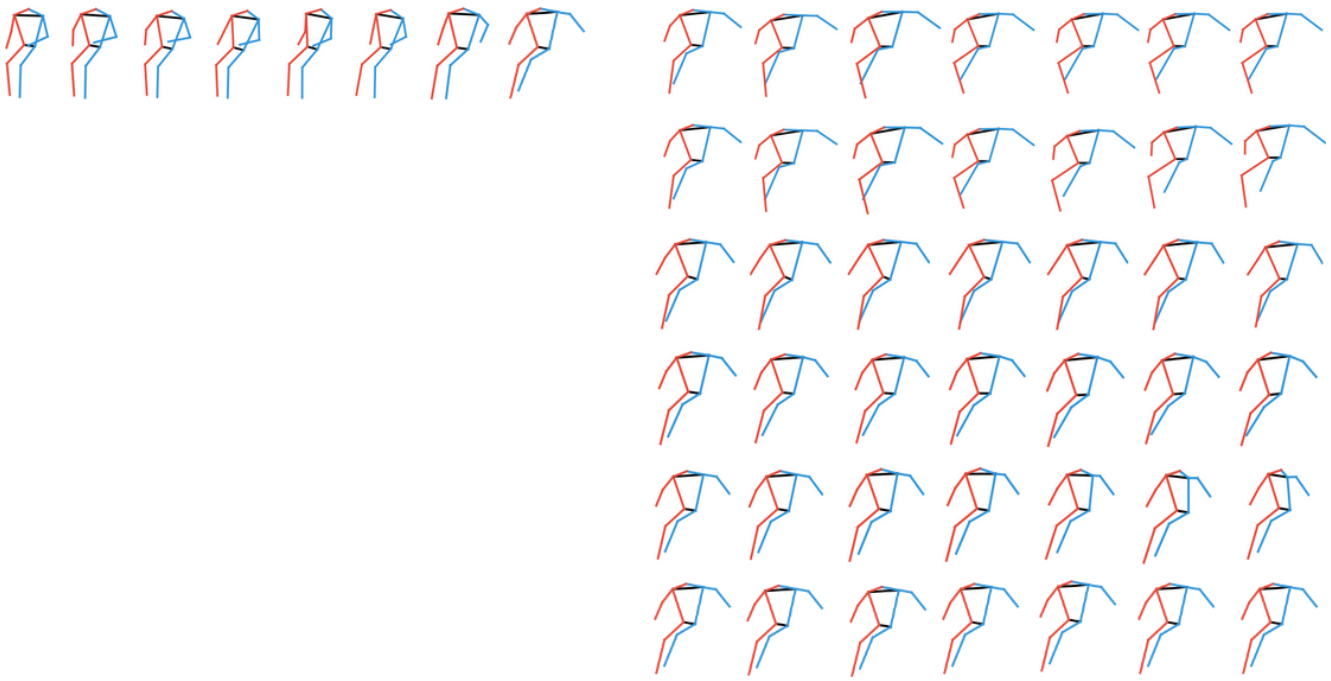
6. Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 754354, the Swiss National Science Foundation under the Grant 2OOO21-L92326, the EPFL Interdisciplinary Seed Fund.

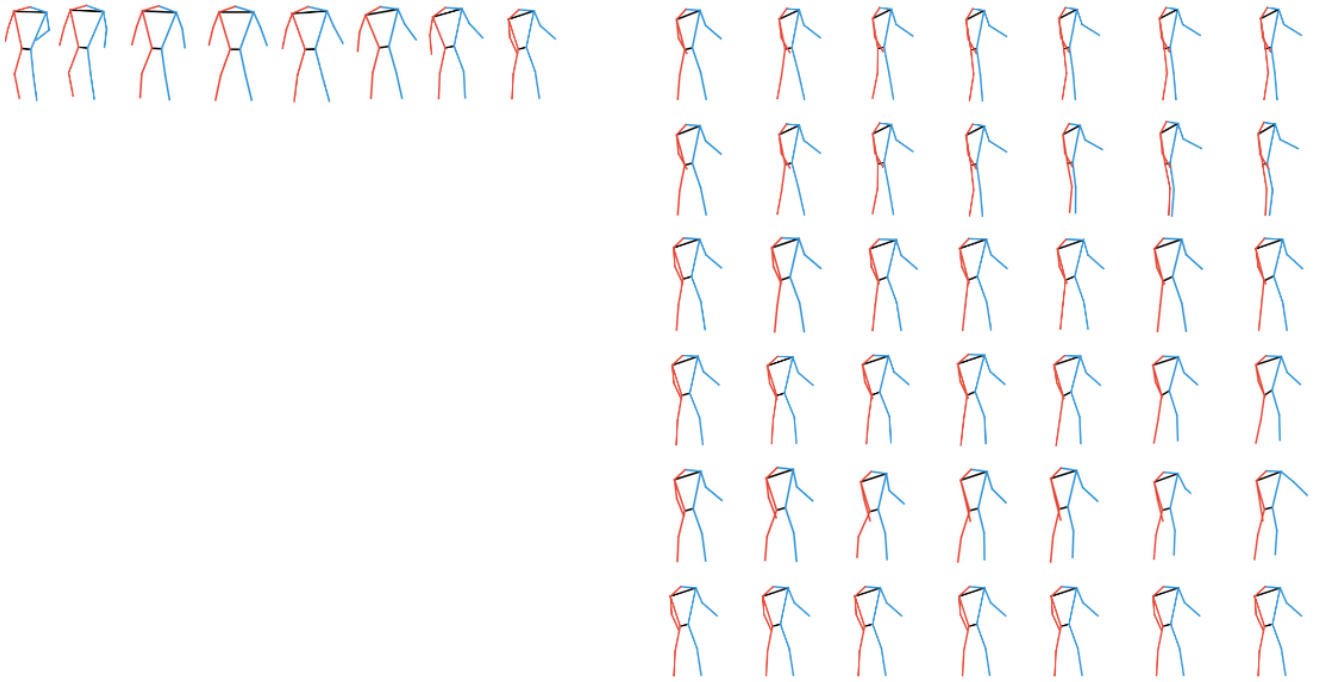
References

[1] Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezatofighi. Socially and contextually aware human motion and pose forecasting. *IEEE Robotics and Automation Letters*, 5(4):6033–6040, 2020. 2, 3, 5, 6

[2] Vida Adeli, Mahsa Ehsanpour, Ian Reid, Juan Carlos Niebles, Silvio Savarese, Ehsan Adeli, and Hamid Rezatofighi. Tripod: Human trajectory and pose dynamics



(a)



(b)

Figure 4: The observed (left) and the predicted (right) pose for two different scenarios in a and b. The rows correspond to the DeRPoF, w/o Early stop, w/o Decoupling, w/o VAE, Decoupling and two baselines Nearest neighbour, and Zero velocity from top to bottom. Only the pose of every other frame is shown.

forecasting in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 4, 5, 6

[3] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, 2016. 1, 2, 5, 6

[4] Alexandre Alahi, Vignesh Ramanathan, and Li Fei-Fei. Socially-aware large-scale crowd forecasting. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2203–2210, 2014. 2

[5] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5223–5232, 2020. 1, 2

[6] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 5167–5176, 2018. 4

[7] Mohammadhossein Bahari, Nejjar Ismail, and Alexandre Alahi. Injecting knowledge in data-driven vehicle trajectory predictors. *Transportation Research Part C*, 2021. 2

[8] Smail Bouhsain, Saeed Saadatnejad, and Alexandre Alahi. Pedestrian intention prediction: A multi-task perspective. In *European Association for Research in Transportation (hEART)*, 2020. 2, 3

[9] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision*, pages 387–404. Springer, 2020. 2

[10] Marco Capogrosso, Fabien B. Wagner, Jerome Gandar, Eduardo Martin Moraud, Nikolaus Wenger, Tomislav Milekovic, Polina Shkorbatova, Natalia Pavlova, Pavel Musienko, Erwan Bezard, Jocelyne Bloch, and Grégoire Courtine. Configuration of electrical spinal cord stimulation through real-time processing of gait kinematics. *Nature Protocols*, 13(9):2031–2061, Sept. 2018. 1

[11] Yu-Wei Chao, Jimei Yang, Brian Price, Scott Cohen, and Jia Deng. Forecasting human dynamics from static images. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 548–556, 2017. 2

[12] Changan Chen, Yuejiang Liu, Sven Kreiss, and Alexandre Alahi. Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning. In *International Conference on Robotics and Automation (ICRA)*, 2019. 1

[13] Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1423–1432. IEEE, 2019. 2

[14] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6992–7001, 2020. 2

[15] Pasquale Coscia, Francesco Castaldo, Francesco AN Palmieri, Alexandre Alahi, Silvio Savarese, and Lamberto Ballan. Long-term path prediction in urban scenarios using circular distributions. *Journal on Image and Vision Computing (JIVC)*, 2018. 2

[16] Qiongjie Cui and Huaijiang Sun. Towards accurate 3d human motion prediction from incomplete observations. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4801–4810, June 2021. 2

[17] Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection. *Neural networks*, 108:466–478, 2018. 2

[18] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4346–4354, 2015. 2, 5

[19] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. In *International Conference on 3D Vision (3DV)*, pages 458–466. IEEE, 2017. 2

[20] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2255–2264, 2018. 1, 2, 5, 6

[21] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995. 2

[22] Joey Hong, Benjamin Sapp, and James Philbin. Rules of the road: Predicting driving behavior with a convolutional model of semantic interactions. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8454–8462, 2019. 2

[23] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6272–6281, 2019. 2, 5, 6

[24] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5308–5317, 2016. 2

[25] Łukasz Kidziński, Bryan Yang, Jennifer L Hicks, Apoorva Rajagopal, Scott L Delp, and Michael H Schwartz. Deep neural networks enable quantitative movement analysis using single-camera videos. *Nature communications*, 11(1):1–10, 2020. 1

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[27] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[28] Parth Kothari, Sven Kreiss, and Alexandre Alahi. Human trajectory forecasting: A deep learning perspective. *IEEE Transactions on Intelligent Transportation Systems*, 2021. 1, 2

[29] Parth Kothari, Brian Sifringer, and Alexandre Alahi. Interpretable social anchors for human trajectory forecasting in crowds. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[30] Parth Ashit Kothari and Alexandre Alahi. Adversarial loss for human trajectory prediction. In *European Association for Research in Transportation (hEART)*, 2019. 2

[31] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 336–345, 2017. 2

[32] Yuejiang Liu, Qi Yan, and Alexandre Alahi. Social nce: Contrastive learning of socially-aware motion representations. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[33] Matthias Luber, Johannes A Stork, Gian Diego Tipaldi, and Kai O Arras. People tracking with human motion predictions from social forces. In *International Conference on Robotics and Automation*, pages 464–469. IEEE, 2010. 2

[34] Karttikeya Mangalam, Ehsan Adeli, Kuan-Hui Lee, Adrien Gaidon, and Juan Carlos Niebles. Disentangling human dynamics for pedestrian locomotion forecasting with noisy supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2784–2793, 2020. 1, 2, 5

[35] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*, pages 474–489. Springer, 2020. 2, 5, 6

[36] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2891–2900, 2017. 1, 2, 3, 5, 6

[37] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 935–942. IEEE, 2009. 2

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 5

[39] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *International Conference on Computer Vision (ICCV)*, pages 261–268. IEEE, 2009. 2

[40] Haziq Razali and Alexandre Alahi. Pedestrian intention prediction: A convolutional bottom-up approach. *Transportation Research Part C*, 2021. 1

[41] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European conference on computer vision*, pages 549–565. Springer, 2016. 2

[42] Christoph Rösmann, Malte Oeljeklaus, Frank Hoffmann, and Torsten Bertram. Online trajectory prediction and planning for social robot navigation. In *International Conference on Advanced Intelligent Mechatronics (AIM)*, pages 1255–1260. IEEE, 2017. 1

[43] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018. 4

[44] Fabien B. Wagner, Jean-Baptiste Mignardot, Camille G. Le Goff-Mignardot, Robin Demesmaeker, Salif Komi, Marco Capogrosso, Andreas Rowald, Ismael Seáñez, Miroslav Caban, Elvira Pirondini, Molywan Vat, Laura A. McCracken, Roman Heimgartner, Isabelle Fodor, Anne Watrin, Perrine Seguin, Edoardo Paoles, Katrien Van Den Keybus, Grégoire Eberle, Brigitte Schurch, Etienne Pralong, Fabio Becce, John Prior, Nicholas Buse, Rik Buschman, Esra Neufeld, Niels Kuster, Stefano Carda, Joachim von Zitzewitz, Vincent Delattre, Tim Denison, Hendrik Lambert, Karen Minassian, Jocelyne Bloch, and Grégoire Courtine. Targeted neurotechnology restores walking in humans with spinal cord injury. *Nature*, 563(7729):65–71, Nov. 2018. 1

[45] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 3332–3341, 2017. 2

[46] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9401–9411, 2021. 2

[47] Nikolaus Wenger, Eduardo Martin Moraud, Jerome Gandar, Pavel Musienko, Marco Capogrosso, Laetitia Baud, Camille G Le Goff, Quentin Barraud, Natalia Pavlova, Nadia Dominici, Ivan R Minev, Leonie Asboth, Arthur Hirsch, Simone Duis, Julie Kreider, Andrea Mortera, Oliver Haverbeck, Silvio Kraus, Felix Schmitz, Jack DiGiovanna, Rubia van den Brand, Jocelyne Bloch, Peter Detemple, Stéphanie P Lacour, Erwan Bézard, Silvestro Micera, and Grégoire Courtine. Spatiotemporal neuromodulation therapies engaging muscle synergies improve motor control after spinal cord injury. *Nature Medicine*, 22:138, Jan. 2016. 1

[48] Kota Yamaguchi, Alexander C Berg, Luis E Ortiz, and Tamara L Berg. Who are you with and where are you going? In *CVPR 2011*, pages 1345–1352. IEEE, 2011. 2

[49] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 265–281, 2018. 2

[50] Jason Y Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. Predicting 3d human dynamics from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7114–7123, 2019. 5